Improving Computational Complexity in Statistical Models with Local Curvature Information

Pedram Akbarian[*1], Tongzheng Ren[*2], Jiacheng Zhuo[2], Sujay Sanghavi[1], Nhat Ho[3]

[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science, [3]Department of Statistics and Data Sciences, The University of Texas at Austin

INTRODUCTION

Problem Setup. We are interested in solving the following optimization problem

$$\widehat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} f_n(\theta), \tag{1}$$

where f_n denote as the *sample* loss function. Moreover, we define the *population* version of the optimization problem (1):

$$\theta^\star \in \arg\min_{\theta \in \mathbb{R}^d} f(\theta) := \mathbb{E}[f_n(\theta)], \tag{2}$$

where f denote as the *population* loss function.

Motivation. The statistical and computational complexity of fixed-step size gradient descent are determined by the singularity of $\nabla^2 f(\theta^\star)$:

	Iterations for convergence	Statistical rate	Computational complexity
Non-singular	$\mathcal{O}(\log(n))$	$\mathcal{O}\left(n^{-1/2}\right)$	$\mathcal{O}(n)$
Singular	$\mathcal{O}\left(n^{\frac{\alpha}{2(\alpha-\gamma)}}\right)$	$\mathcal{O}\left(n^{\frac{-1}{2(\alpha-\gamma)}}\right)$	$\mathcal{O}\left(n^{1+\frac{\alpha}{2(\alpha-\gamma)}}\right)$

Table 1: Suboptimality of gradient descent iterates for singular models.

To overcome the suboptimal computational complexity of the GD algorithm, we consider the utilization of the *local curvature information*. In this work, we specifically address the following question:

Is there a method that achieves a balance between computational efficiency and provable statistical optimality at a reasonable per-iteration computational cost?

We explore this inquiry and demonstrate that the normalized gradient descent (NormGD) algorithm can attain both statistical optimality and computational efficiency.

Contributions.

1. **General Theory.** We study the computational and statistical complexity of NormGD iterates when the population loss function is homogeneous in all directions, and the stability of first-order and second-order information holds.
2. **Examples.** We illustrate the general theory for the statistical guarantee of NormGD under two popular statistical models: (1) Generalized Linear Models (GLM) and (2) Gaussian Mixture Models (GMM).

MAIN RESULTS

Normalized Gradient Descent (NormGD). The iterative steps of Normalized Gradient Descent (NormGD) for the sample and population loss functions are given by $\theta_n^{t+1} := F_n^{\text{NGD}}(\theta_n^t)$ and $\theta^{t+1} := F^{\text{NGD}}(\theta^t)$, respectively. The definitions of the NormGD operators for the sample and population cases are as follows:

$$F_n^{\text{NGD}}(\theta_n^t) := \theta_n^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f_n(\theta_n^t))}\nabla f(\theta_n^t) \qquad \text{(Sample iterate)}$$

$$F^{\text{NGD}}(\theta^t) := \theta^t - \frac{\eta}{\lambda_{\max}(\nabla^2 f(\theta^t))}\nabla f(\theta^t) \qquad \text{(Population iterate)}$$

Assumption 1. (Homogeneous Property)

Given the constant $\alpha > 0$ and the radius $r > 0$, for all $\theta \in \mathbb{B}(\theta^\star, r)$ we have

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq c_1 \|\theta - \theta^\star\|^\alpha,$$
$$\lambda_{\max}(\nabla^2 f(\theta)) \leq c_2 \|\theta - \theta^\star\|^\alpha,$$

where $c_1 > 0$ and $c_2 > 0$ are some universal constants depending on r.

Assumption 2. (Stability of Second-order Information)

For a given parameter $\gamma \geq 0$, there exist a noise function $\varepsilon : \mathbb{N} \times (0, 1] \to \mathbb{R}^+$, universal constant $c_3 > 0$, and some positive parameter $\rho > 0$ such that

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\nabla^2 f_n(\theta) - \nabla^2 f(\theta)\|_{\text{op}} \leq c_3 r^\gamma \varepsilon(n, \delta),$$

for all $r \in (0, \rho)$ with probability $1 - \delta$.

Theorem (Informal)

Assume that assumptions (1) and (2) hold with $\alpha \geq \gamma + 1$. Then, there exist universal constants C_1, C_2 such that with probability $1 - \delta$, for $t \geq C_1 \log(1/\varepsilon(n, \delta))$, the following holds:

$$\min_{k \in \{0,1,\cdots,t\}} \|\theta_n^k - \theta^\star\| \leq C_2 \cdot \varepsilon(n, \delta)^{\frac{1}{\alpha-\gamma}}.$$

Generalized Linear Model (GLM). Let $\{(Y_i, X_i)\}_{i=1}^n$ satisfy

$$Y_i = g(X_i^\top \theta^\star) + \varepsilon_i. \qquad \forall i \in [n] \tag{3}$$

Here, $g : \mathbb{R} \to \mathbb{R}$ is a given link function, $\theta^\star$ is a true but unknown parameter, and ε_i are i.i.d. noises from $\mathcal{N}(0, \sigma^2)$.

Least-square loss. We estimate the true parameter $\theta^\star$ via minimizing the least-square loss function:

$$\mathcal{L}_n(\theta) := \frac{1}{2n}\sum_{i=1}^n (Y_i - (X_i^\top \theta)^p)^2, \qquad \text{(Sample loss)}$$

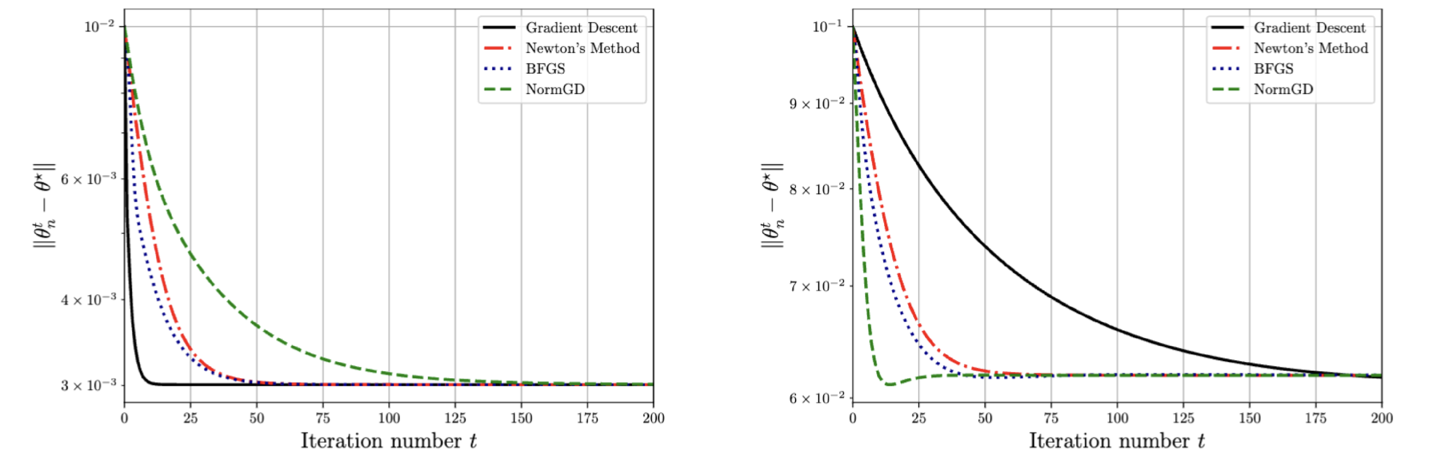
$$\mathcal{L}(\theta) := \frac{1}{2}\mathbb{E}[(Y - (X^\top \theta)^p)^2]. \qquad \text{(Population loss)}$$

Table 2: Overview of Results for GLM with Link Function $g(r) = r^p$ in low SNR regime with $\theta^\star = 0$.

Algorithm	Iterations for convergence	Statistical error on convergence	Computational complexity
Gradient Descent	$(n/d)^{\frac{p-1}{p}}$	$(d/n)^{\frac{1}{2p}}$	$n^{\frac{2p-1}{p}} d^{\frac{1}{p}}$
Newton's Method	$\log(n/d)$	$(d/n)^{\frac{1}{2p}^*}$	$(nd + d^3)\log(n/d)$
BFGS	$\log(n/d)$	$(d/n)^{\frac{1}{2p+2}^*}$	$(nd + d^2)\log(n/d)$
NormGD (Ours)	$\log(n/d)$	$(d/n)^{\frac{1}{2p}}$	$(nd + d^2)\log(n/d)$

NUMERICAL EXPERIMENTS

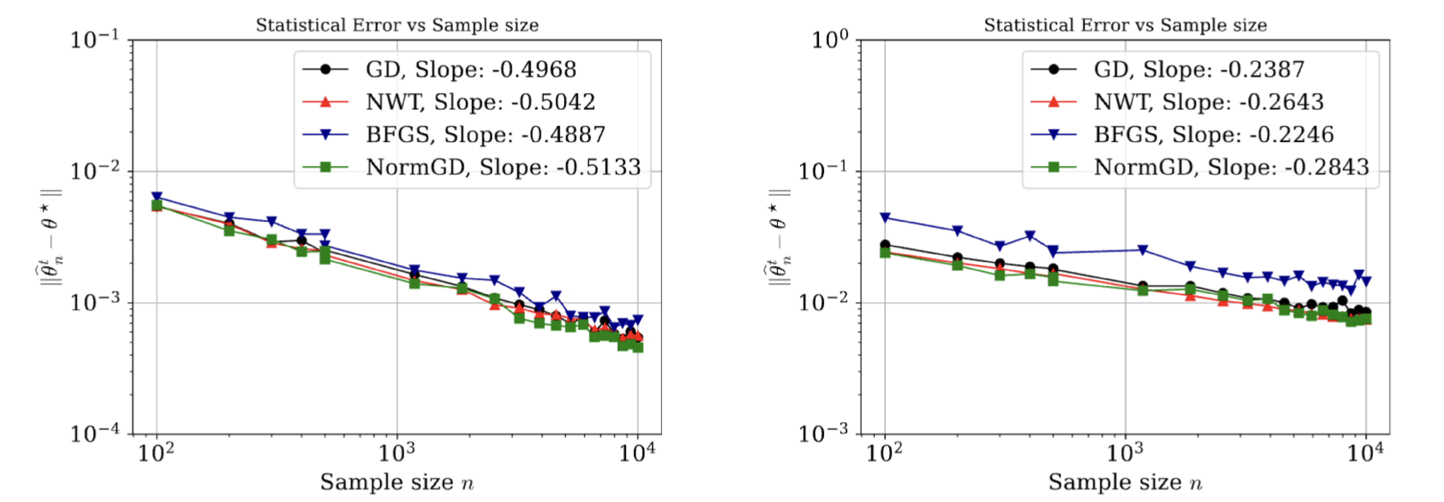
Figure 1: **Left**: All methods converge linearly in the high signal-to-noise setting; **Right**: all second-order methods converge linearly in the low signal-to-noise setting while GD converges sub-linearly.



(a) Convergence rate of sample iterates in High SNR setting with $n = 1000$

(b) Convergence rate of sample iterates in Low SNR setting with $n = 1000$

Figure 2: **Left**: (High SNR) The statistical error of all methods roughly scales with $n^{-0.5}$; **Right**: (Low SNR) the statistical error roughly scales with $n^{-0.25}$ for all methods.



(a) Statistical rate of sample iterates in High SNR setting with $d = 2$

(b) Statistical rate of sample iterates in Low SNR setting with $d = 2$